Big Data Analytics

Attività – Data Analytics

Data processing e exploratory data anlytics su Data set provenienti da più sorgenti

Obiettivo in breve: L'attività consiste nello sviluppare un progetto di Data Analytics in un ambito a proprio piacere e di proprio interesse finalizzato allo storytelling ovvero a trovare risposta a più quesiti di ricerca.

Quanti e quali dataset? Il progetto deve partire da due o più dataset di vari formati non necessariamente distinti (csv, json,ecc.) provenienti da *almeno due sorgenti distinte*. Diversi siti Web consentono di scaricare dataset pubblici come ad esempio open data, è possibile effettuare lo scraping di dati pubblicati in pagine Web, infine sul sito di Kaggle all'indirizzo https://www.kaggle.com/datasets sono disponibili diversi data set. Le pagine da dove è possibile scaricare i dataset ne forniscono una descrizione e, in alcuni casi, anche suggerimenti di quesiti che potrebbero essere indagati usando il dataset stesso.

Descrizione: L'attività da svolgere consiste nel:

- Scegliere due o più dataset provenienti da due o più sorgenti. Il dataset finale deve essere costituito almeno da due file.
- Usando PANDAS implementare le operazioni di data processing necessarie (principalmente join e selezioni) per mettere in collegamento i dataset e per preparare i dati al passo successivo
- Usando pacchetti Python quali Pandas, scipy, matplotlib e sciborn implementare attività di data cleaning, exploratory data analysis estraendo dati statistici e di visualizzazione dei risultati attraverso il quale sia possibile "raccontare qualcosa sui dati" (storytelling), eventualmente partendo da dei quesiti di ricerca. L'uso dei pacchetti non deve necessariamente essere limitato alle istruzioni viste a lezione. Le documentazioni dei pacchetti stessi e i volumi messi a disposizione su Dolly fornisco spunti d'uso interessanti!!
- Produrre un notebook Jupyter (https://jupyter.org/) che contenga:
    - Una introduzione all'argomento scelto, alle sorgenti dati e agli obiettivi del progetto specificando eventualmente i quesiti di ricerca
    - Una sezione per ogni fase del progetto di data analytics

Oltre ad un esempio di progetto degli scorsi scaricabile da Moodle, altri esempi di progetti di data analytics sono i seguenti:
https://jiglesia3.github.io/
https://andresgogo.github.io/
https://megcren.github.io/ - esempio di progetto ampio ma poco commentato
https://meteosr.github.io/

Consegna: Upload del notebook al corrispondente link nella pagina Moodle del corso.

Valutazione: Le attività verranno valutate sulla base dei seguenti criteri:
1. Motivazione. L'elaborato fa credere al lettore che l'argomento sia rilevante o importante (i) in generale e (ii) rispetto alla scienza dei dati?
2. Comprensione. Dopo aver letto l'elaborato, un lettore non informato si sente informato il tema? Un lettore che già conosce l'argomento ha l'impressione di aver imparato di più?
3. Storytelling. La parte in prosa dell'elaborato è convincente?
4. Codice. Il codice è ben scritto, ben documentato, riproducibile e aiuta il lettore a capire? Fornisce buoni esempi di tecniche specifiche?

Scadenza per premio partecipazione: **7/12/2022**